Multilingual Recurrent Neural Networks with Residual Learning for Low-Resource Speech Recognition

Shiyu Zhou[1,2], *Yuanyuan Zhao*[1,2], *Shuang Xu*[1], *Bo Xu*[1]

[1]Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
{zhoushiyu2013, yyzhao5231, shuang.xu, xubo}@ia.ac.cn

Abstract

The shared-hidden-layer multilingual deep neural network (SHL-MDNN), in which the hidden layers of feed-forward deep neural network (DNN) are shared across multiple languages while the softmax layers are language dependent, has been shown to be effective on acoustic modeling of multilingual low-resource speech recognition. In this paper, we propose that the shared-hidden-layer with Long Short-Term Memory (LSTM) recurrent neural networks can achieve further performance improvement considering LSTM has outperformed DNN as the acoustic model of automatic speech recognition (ASR). Moreover, we reveal that shared-hidden-layer multilingual LSTM (SHL-MLSTM) with residual learning can yield additional moderate but consistent gain from multilingual tasks given the fact that residual learning can allievate the degradation problem of deep LSTMs. Experimental results demonstrate that SHL-MLSTM can relatively reduce word error rate (WER) by 2.1-6.8% over SHL-MDNN trained using six languages and 2.6-7.3% over monolingual LSTM trained using the language specific data on CALLHOME datasets. Additional WER reduction, about relatively 2% over SHL-MLSTM, can be obtained through residual learning on CALLHOME datasets, which demonstrates residual learning is useful for SHL-MLSTM on multilingual low-resource ASR.

Index Terms: LSTM, multilingual speech recognition, low-resource, residual learning, shared-hidden-layer

1. Introduction

Nowadays experts have shown significant interest in the area of multilingual acoustic modeling by the context-dependent deep neural network hidden Markov models (CD-DNN-HMM) [1, 2]. Many recent studies have shown that multilingual data can be used to improve the monolingual ASR performance in CD-DNN-HMM [3, 4]. The hidden layers of DNN in CD-DNN-HMM can be thought of complicated feature transformation through multiple layers of nonlinearity, which can be used to extract universal feature transformation from multilingual datasets [4]. Among the CD-DNN-HMM based approaches, the architecture of SHL-MDNN [4], in which the hidden layers are shared across multiple languages while the softmax layers are language dependent, is a significant progress in the area of multilingual ASR. The shared hidden layers and the language dependent softmax layers of SHL-MDNN are optimized jointly by multilingual datasets, which can be considered as a universal feature transformation that works well for multilingual ASR [4].

The work was supported by 973 Program in China, grant No. 2013CB329302.

LSTM recurrent neural networks [5, 6] contain special units called memory blocks to store the temporal state of the networks and multiplicative units called gates to control the flow of information. Because LSTM recurrent neural networks have powerful ability of modeling long-span dependency, several previous researches [6, 7, 8] have shown that it has achieved state-of-the-art performance on Large Vocabulary Continuous Speech Recognition (LVCSR) tasks compared with other acoustic models such as Gaussian mixture models (GMMs) and DNNs.

Because LSTM recurrent neural networks have outperformed DNN as the acoustic model of ASR, we argue that the shared-hidden-layer using LSTM can outperform that using DNN in multilingual low-resource ASR. Although the difference of languages might influence the memory blocks of LSTM when training, we suppose the shared-hidden-layer using LSTM can still benefit and learn more invariant universal feature transformation from multiple languages compared with that using DNN. Experimental results confirm our hypothesis that SHL-MLSTM can relatively reduce WER by 2.1-6.8% over SHL-MDNN trained using 6 languages and 2.6-7.3% over monolingual LSTM trained using the language specific data on CALLHOME datasets. Moreover, the number of languages is investigated and experimental results reveal that the WER almost decreases as the number of languages increases. The results indicate SHL-MLSTM can benefit from multiple languages.

As depicted in paper [9], deeper neural networks are more difficult to train because of the degradation problem. To address this problem, residual learning has been proposed [9] in which added layers are identity map connecting shallower layer to higher layer. The extra identity shortcut connections add neither extra parameter nor computational complexity, but work very well in deep convolutional neural networks (CNNs). Zhao et al. [10] further experimentally investigated LSTM with residual learning to improve the performance of ASR. So SHL-MLSTM with residual learning is investigated in this paper to alleviate the degradation problem as the shared-hidden-layers increase. Experimental results reveal that SHL-MLSTM with residual learning can obtain about relatively 2% WER reduction over SHL-MLSTM on CALLHOME datasets which prove residual learning is effective for SHL-MLSTM.

The rest of the paper is organized as follows. After an overview of the related work in Section 2, Section 3 describes the architecture of SHL-MLSTM with residual learning in detail. we then show experimental results in Section 4 and conclude this work in Section 5.

2. Related work

Multilingual speech recognition has been studied [4, 11, 12] for a long time. It has been shown that the knowledge learned from

multiple languages can be used to improve the performance of each individual language. In the conventional GMM-HMM multilingual speech recognition, Subspace Gaussian Mixture Model (SGMM) has been studied [11] in which majority of parameters are shared across the states and trained by data from multiple languages. Since the modeling capability of discriminative DNNs is significantly more powerful than that of the generative Gaussian mixtures model, Huang et al. [4] proposed SHL-MDNN, in which the hidden layers with DNN are shared across multiple languages while the softmax layers are language dependent. For low-resource ASR configurations, low-rank factorization of weight matrices via Singular Value Decomposition [1] was investigated to reduce the number of parameters of multilingual DNN.

It is more difficult to train deep neural networks along with the increase of DNN depth because of the degradation problem. He et al. [9] proposed residual learning [9] to alleviate it in deep CNNs. Residual networks can be seen as a collection of many paths of differing length and behave like ensembles of relatively shallow networks [13]. Zhao et al. [10] further investigated the deep multidimensional residual networks with LSTM to alleviate the degradation problem in deep RNNs. In the spatial dimension, shortcut connections was introduced to RNNs, along which the information can flow across several layers without attenuation. In the temporal dimension, the input sequence was split into several parallel sub-sequences to ensure information flowing across the time axis unimpededly.

3. SHL-MLSTM with residual learning

This section illustrates the architecture of SHL-MLSTM with residual learning. First, SHL-MLSTM will be described, and then SHL-MLSTM with residual learning will be detailed in the following subsection.

3.1. SHL-MLSTM

A LSTM network computes a mapping from an input sequence x = (x_1, ..., x_T) to an output sequence y = (y_1, ..., y_T) by calculating the network unit activations using the following equations iteratively from t = 1 to T [6]:

$$i_t = \sigma\left(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i\right) \quad (1)$$

$$f_t = \sigma\left(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f\right) \quad (2)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g\left(W_{cx}x_t + W_{cm}m_{t-1} + b_c\right) \quad (3)$$

$$o_t = \sigma\left(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o\right) \quad (4)$$

$$m_t = o_t \odot h\left(c_t\right) \quad (5)$$

$$y_t = W_{ym}m_t + b_y \quad (6)$$

where the *W* terms denote weight matrices (e.g. W_{ix} is the matrix of weights from the input gate to the input), the *b* terms denote bias vectors (b_i is the input gate bias vector), σ is the logistic sigmoid function, and *i*, *f*, *o* and *c* are respectively the input gate, forget gate, output gate and cell activation vectors, all of which are the same size as the cell output activation vector *m*, $\odot$ is the element-wise product of the vectors and *g* and *h* are the cell input and cell output activation functions, generally *tanh*.

The proposed architecture of SHL-MLSTM is illustrated in Figure 1 (without dotted shortcut connections), in which the hidden layers with LSTM instead of DNN are shared across multiple languages while the softmax layers same as SHL-MDNN are language dependent. The shared-hidden-layers with

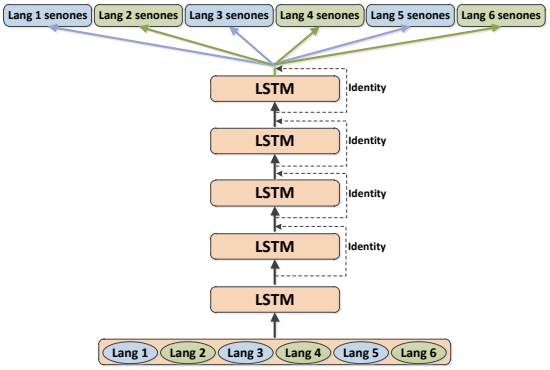

Figure 1: *The architecture of SHL-MLSTM with(/without) residual learning.*

LSTM can be considered as a universal feature transformation of multiple languages and the language dependent softmax layers estimate the posterior probabilities of the senones (tied triphone states) specific to that language. To train SHL-MLSTM successfully, the greedy layer-wise supervised pre-training algorithm is adopted [14]. During pre-training, all languages dependent softmax layers are used as supervisors and the training utterances are randomized [4] before feeding into SHL-MLSTM. The slightly adjusted backpropagation (BP) algorithm is adopted [4] in the fine-tuning procedure. During training, when one language training sample is filled in the SHL-MLSTM, only the shared hidden layers and the language-specific softmax layer are updated. Other languages' softmax layers are kept intact. When training is finished, SHL-MLSTM can work as a multiple language speech recognizer for all the languages used in the training procedure.

3.2. SHL-MLSTM with residual learning

A residual learning framework [9] lets stacked layers fit a residual mapping explicitly, instead of hoping these layers directly fit a desired underlying mapping. Formally, denoting the desired underlying mapping as H(x), we let the stacked nonlinear layers fit another mapping of F(x) := H(x) - x, and then the original mapping is recast into F(x) + x. The formulation of F(x) + x can be realized by identity shortcut connections.

The major part of LSTM with residual learning is still the same as plain LSTM networks except that the inputs of the *l* + *2* layer $\tilde{x}_t$ are made up of two parts: the outputs of the *l* layer and the *l* + *1* layer. It formulates as (7), and the formula of LSTM with residual learning can be represented as (8)-(13) [10].

$$\tilde{x}_t = y_t^l \oplus y_t^{l+1} \quad (7)$$

$$i_t = \sigma\left(W_{ix}\tilde{x}_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i\right) \quad (8)$$

$$f_t = \sigma\left(W_{fx}\tilde{x}_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f\right) \quad (9)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g\left(W_{cx}\tilde{x}_t + W_{cm}m_{t-1} + b_c\right) \quad (10)$$

$$o_t = \sigma\left(W_{ox}\tilde{x}_t + W_{om}m_{t-1} + W_{oc}c_t + b_o\right) \quad (11)$$

$$m_t = o_t \odot h\left(c_t\right) \quad (12)$$

$$y_t = W_{ym}m_t + b_y \quad (13)$$

There are two ways to deal with the operation $\oplus$ in formulation (7): a) y_t^l concatenates with y_t^{l+1}; b) y_t^l sums with y_t^{l+1}.

We choose the sum function as operation $\oplus$ in order to keep parameters consistent with corresponding plain LSTM.

The architecture of SHL-MLSTM with residual learning is the same as SHL-MLSTM except appending an identity connection between LSTMs. Each shared-hidden-layer consists of a residual LSTM module f_i and an identity connection bypassing f_i. It is depicted in Figure 1 (with dotted shortcut connections), in which shortcut connections with identity mapping are appended between LSTMs without extra parameters. The identity mapping between layers can allow data to flow directly and carry gradient through deep LSTMs alleviating the degradation problem, so SHL-MLSTM with residual learning can learn better universal feature transformation from multiple languages than plain SHL-MLSTM.

4. Experiment

4.1. Datasets and setup

The datasets in the paper come from CALLHOME corpora collected by Linguistic Data Consortium (LDC), which used in [3]. The following six languages are used: Mandarin (MA), English (EN), Japanese (JA), Arabic (AR), German (GE) and Spanish (SP). The detailed information is listed below in Table 1.

Table 1: *Training and testing corpora information for the six languages.*

Language	Training hours	Testing hours	#phn	Lexicon size	PPL
Mandarin (MA)	15.6	1.5	38	$44K$	167
English (EN)	14.9	1.8	43	$91K$	108
Japanese (JA)	15.1	2.1	34	$135K$	73
Arabic (AR)	13.6	1.4	41	$51K$	195
German (GE)	14.7	1.8	46	$319K$	119
Spanish (SP)	16.5	1.9	29	$46K$	135

4.2. Baseline systems

All experiments are conducted using 42 dimension features which consist of 13-dimensional PLP and smoothed F0 appended with the first and second order derivatives. The features are normalized via mean subtraction and variance normalization on the speaker basis. For each language, standard tied-state cross-word triphone GMM-HMMs are first trained with maximum likelihood estimation as [3].

Our baseline systems are monolingual DNN, monolingual LSTM and SHL-MDNN. DNN based models contain 5 hidden layers with 2048 units in each layer. LSTM based models contain 5 hidden layers with 800 LSTM cells projected to 512 units in each layer. The multilingual models have 6 softmax layers (one layer per each language). The RBM-based unsupervised pre-training algorithm [15] is performed on DNN based models while the greedy layer-wise supervised pre-training algorithm [14] is performed on LSTM based models, followed by the fine-tuning process using stochastic gradient descent with the cross-entropy criterion. In order to ensure training stability, the activations of memory cells of SHL-MLSTM are cliped to range [-50, 50] and the gradients to [-1, 1] separately.

4.3. The results of SHL-MLSTM

The results of SHL-MLSTM are summarized in Table 2. First, we can note that the WER of SHL-MDNN (MA+EN+JA+AR+GE+SP) is comparable with monolingual LSTM, although the WER of monolingual DNN is less than it. It indicates that although DNN is not as good as LSTM from the perspective of sequence modeling capability, it can be compensated by the shared-hidden-layer with DNN from training multiple languages. Second, from the comparison between monolingual LSTM and SHL-MLSTM (MA+EN+JA+AR+GE+SP), we can observe that SHL-MLSTM outperforms the monolingual LSTM with a 2.6-7.3% relative WER reduction across six languages. It indicates that LSTM can be used as the shared-hidden-layer and benefit from training multiple languages. Third, we can observe that SHL-MLSTM (MA+EN+JA+AR+GE+SP) outperforms SHL-MDNN (MA+EN+JA+AR+GE+SP) with about 2.1-6.8% relative WER reduction across six languages. It reveals that LSTM is superior to DNN as the shared-hidden-layer in ASR because of more powerful sequence modeling capability of LSTM.

In addition, we investigate the relationship between the number of languages and WER in SHL-MLSTM on CALLHOME datasets. First, we train SHL-MLSTM with two languages of mixed MA and EN. The results show that the WER of MA from SHL-MLSTM (MA+EN) is better than monolingual LSTM of MA, so is EN. And then we train SHL-MLSTM sequentially increasing the number of languages one by one. The order of added languages is arbitrary without special consideration. From Table 2, it shows that the performance of all languages are almost improved gradually as the number of language increases. The results reveal that SHL-MLSTM can benefit from multiple languages, although there might be a conflict among different languages when training SHL-MLSTM, e.g. the added language AR leads to performance decrease slightly in language MA and EN when training SHL-MLSTM (MA+EN+JA+AR).

4.4. The results of SHL-MLSTM with residual learning

The results of SHL-MLSTM with residual learning are shown in Table 3. Two approaches are used to pre-train SHL-MLSTM with residual learning in the experiments. The first method is to pre-train SHL-MLSTM with residual learning using greedy layer-wise supervised pre-training technique [14] as did in SHL-MLSTM. And the second method is to append the identity map directly to the trained SHL-MLSTM as the pre-trained model. After pre-trained, the same method as SHL-MLSTM is adopted to fine-tune the pre-trained model of SHL-MLSTM with residual learning. From the results of Table 3, we can find that both of these two pre-training approaches are useful and the second seems work slightly better. That is to say, if we have trained a SHL-MLSTM model from multiple languages, it is a good choice to append the identity map directly to the trained SHL-MLSTM and train SHL-MLSTM with residual learning directly so as to both save the pre-training time and obtain better performance. The results of Table 3 demonstrate that SHL-MLSTM-RESIDUAL (SHL-INIT) can get additional relative 1.7-2.0% WER reduction compared with SHL-MLSTM on CALLHOME datasets. It proves SHL-MLSTM using residual learning can benefit from the appended identity mapping, which allows data to flow directly and carry gradient through deep LSTMs.

After SHL-MLSTM-RESIDUAL (SHL-INIT) being trained, the language adaption is explored to improve performance of each language further. The shared-hidden-layers are extracted from SHL-MLSTM-RESIDUAL (SHL-INIT)

Table 2: *Comparison of baseline systems and SHL-MLSTM on CALLHOME datasets in WER (%)*

Model	Languages	#Parameters	MA	EN	JA	AR	GE	SP	Average
DNN	Monolingual	≈21.0M	53.05	50.45	57.52	61.52	59.11	59.77	56.90
LSTM	Monolingual	≈17.8M	50.53	48.16	55.14	59.21	56.61	57.71	54.56
SHL-MDNN	MA+EN+JA+AR+GE+SP	38.0M	50.67	46.77	54.15	58.91	55.94	57.88	54.05
SHL-MLSTM	MA+EN	18.6M	49.33	47.15	–	–	–	–	–
SHL-MLSTM	MA+EN+JA	19.5M	48.13	45.93	51.91	–	–	–	–
SHL-MLSTM	MA+EN+JA+AR	20.3M	48.28	46.03	51.39	57.72	–	–	–
SHL-MLSTM	MA+EN+JA+AR+GE	21.1M	47.72	45.47	51.37	58.15	53.78	–	–
SHL-MLSTM	MA+EN+JA+AR+GE+SP	22.0M	**47.24**	**44.94**	**51.11**	**57.67**	**52.92**	**54.31**	**51.37**

Table 3: *Comparison of SHL-MLSTM and SHL-MLSTM-RESIDUAL on CALLHOME datasets in WER (%)*

Model	MA	EN	JA	AR	GE	SP	Average
SHL-MLSTM	47.24	44.94	51.11	57.67	52.92	54.31	51.37
SHL-MLSTM-RESIDUAL (LAYERWISE)	45.87	43.96	**49.78**	56.88	52.39	53.74	50.44
SHL-MLSTM-RESIDUAL (SHL-INIT)	**45.85**	**43.93**	50.13	**56.47**	**51.75**	**53.38**	**50.25**
SHL-MLSTM-RESIDUAL (SHL-INIT ADAPTION)	**45.75**	**42.76**	**49.46**	**55.78**	**51.26**	**52.63**	**49.61**

Table 4: *Different layers of SHL-MLSTM with(/without) residual learning on CALLHOME datasets in WER (%)*

Model	Layer	MA	EN	JA	AR	GE	SP	Average
SHL-MLSTM	3L	47.25	**44.59**	**50.72**	57.48	52.98	54.63	51.28
	4L	47.28	44.92	**50.72**	**57.30**	52.67	54.71	**51.27**
	5L	47.24	44.94	51.11	57.67	52.92	**54.31**	51.37
	6L	**46.85**	45.16	51.17	57.82	**51.90**	54.37	51.39
	7L	47.31	44.81	51.00	57.94	52.85	54.39	51.39
	8L	47.24	44.96	51.47	57.71	53.18	54.42	51.50
	9L	47.24	45.14	51.35	57.78	52.89	54.95	51.56
SHL-MLSTM-RESIDUAL	3L	45.93	43.59	49.98	56.44	51.76	53.79	50.25
	4L	46.10	43.87	49.88	56.68	51.79	53.81	50.36
	5L	45.85	43.93	50.13	56.47	51.75	53.38	50.25
	6L	45.59	**42.78**	49.52	**55.37**	**51.30**	53.12	**49.61**
	7L	45.61	43.14	**49.48**	55.98	51.67	53.49	49.90
	8L	45.66	42.93	49.55	55.49	51.58	**53.00**	49.70
	9L	**45.27**	43.33	49.59	55.76	51.81	53.25	49.84

and each language specific softmax layer is added on top of it to produce 6 SHL-based monolingual LSTMs. Then, the language adaptation is implemented to retrain each SHL-based monolingual LSTM by using the language specific training data and it achieves a better performance compared with the models without language adaption.

4.5. The depth of shared-hidden-layers

Finally, the depth of shared-hidden-layers (SHLs) is investigated. The results are shown in Table 4. We explore different depth from 3 SHLs to 9 SHLs and observe the change of performance. First, it is shown that SHL-MLSTM with residual learning always outperforms corresponding plain SHL-MLSTM, which indicates residual learning is useful for training deep SHL-MLSTM. Second, the performance of SHL-MLSTM becomes slightly worse as the depth increases. This is probably because it is more difficult to train deep SHL-MLSTM as the depth increases due to the degradation problem. Third, the performance of SHL-MLSTM with residual learning becomes slightly better as the depth increases and the best result is obtained in depth with 6 SHLs. But as the depth increases from 6 SHLs to 9 SHLs, the performance has not obvious improvement. Our interpretation of this phenomenon is that residual learning can alleviate the degradation problem of deep SHL-MLSTM, but as the depth increases, models are not trained fully since there is not enough training data.

5. Conclusions

In this paper, we proposed the SHL-MLSTM architecture, in which the hidden layers with LSTM instead of DNN are shared across multiple languages while the softmax layers same as SHL-MDNN are language dependent. We verified the effectiveness of the proposed SHL-MLSTM through the reduction of WERs compared with SHL-MDNN and monolingual LSTM on CALLHOME datasets. Futhermore, the architecture of SHL-MLSTM with residual learning is investigated and two pre-training approaches are compared. The experimental results demonstrate that SHL-MLSTM with residual learning outperforms plain SHL-MLSTM. Last, the depth of SHLs is studied and the results indicate residual learning is useful for training deep SHL-MLSTM.

6. References

[1] R. Sahraeian and D. Van Compernolle, "A study of rank-constrained multilingual dnns for low-resource asr," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5420–5424.

[2] A. Mohan and R. Rose, "Multi-lingual speech recognition with low-rank multi-task deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4994–4998.

[3] J. Li, X. Wang, B. Xu *et al.*, "An empirical study of multilingual and low-resource spoken term detection using deep neural networks." in *INTERSPEECH*, 2014, pp. 1747–1751.

[4] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[7] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[8] A. Senior, H. Sak, and I. Shafran, "Context dependent phone models for lstm rnn acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4585–4589.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[10] Y. Zhao, S. Xu, and B. Xu, "Multidimensional residual learning based on recurrent neural networks for acoustic modeling," *Interspeech 2016*, pp. 3419–3423, 2016.

[11] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4334–4337.

[12] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7639–7643.

[13] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 550–558.

[14] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[15] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.